Machine Learning Assignment

PROJECT REPORT

<TEAM ID : 12 >

Reddit Post Popularity Regression

Name	SRN
Deesha C	PES2UG23CS165
Chitra Madarakhandi	PES2UG23CS157

Problem Statement

In this project, the goal is to predict the popularity of Reddit posts measured by upvotes or scores using regression models. Students will preprocess text and metadata features such as post length, posting time, subreddit, and user information. Model selection may include linear regression, random forests, or gradient boosting regressors. Performance metrics include mean squared error and mean absolute error. Feature importance analysis will highlight factors contributing the most to post popularity. The dataset includes Reddit posts with labeled popularity, simulating real social media analytics challenges.

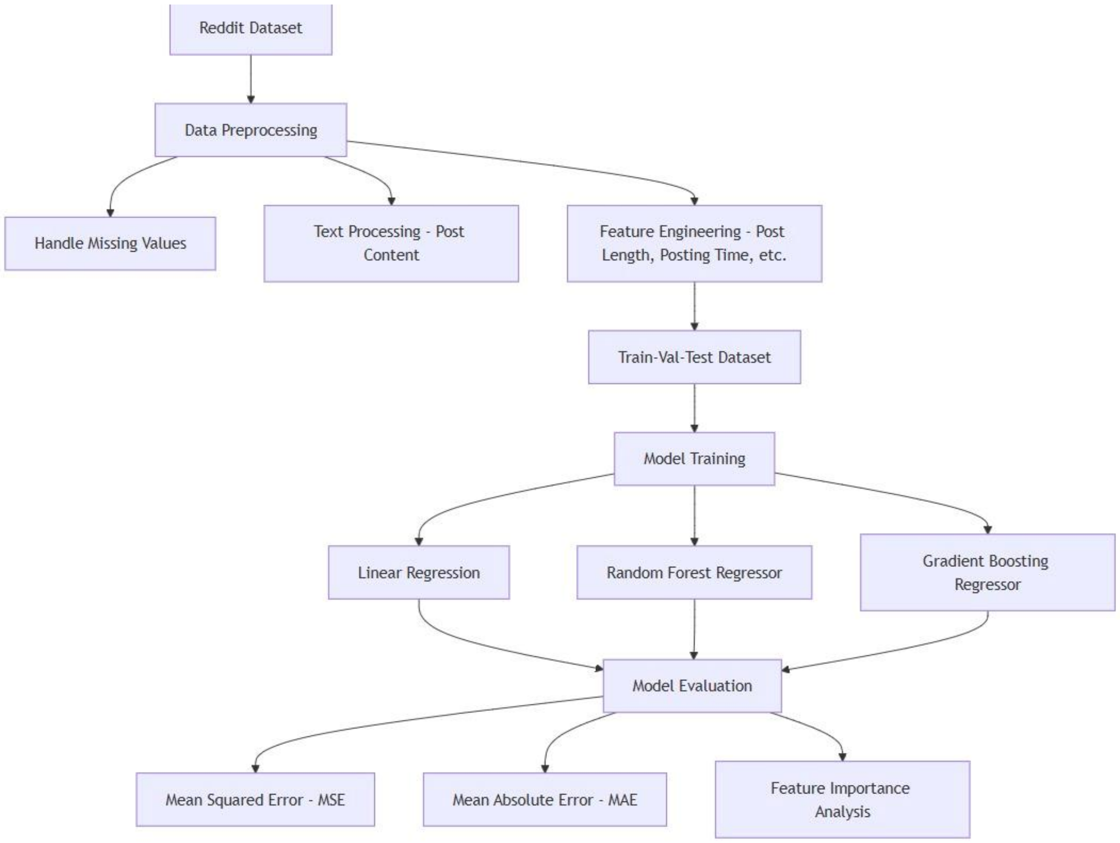
Objective / Aim

The objective of this project is to develop machine learning regression models that predict the popularity of Reddit posts, measured by their scores or upvotes, using both textual content and metadata features. The project involves preprocessing post text, posting time, subreddit, and user-related data to create meaningful features, and then training models such as linear regression, random forest, and gradient boosting to capture the complex relationships affecting popularity. Evaluation metrics like mean squared error and mean absolute error are used to assess model performance, while feature importance analysis provides insight into the key factors driving post engagement. Overall, the study aims to simulate real social media analytics challenges by forecasting which posts are likely to become popular, helping understand user interaction and content dynamics on Reddit.

Dataset Details

- **Source:** Kaggle
- **Size:** Train: (214390, 21) Validation: (75031, 21) Test: (75022, 21)
- **Key Features:**
  1.name: A unique Reddit identifier combining type and ID.
  2.link_id: The unique ID of the parent Reddit post to which the comment belongs.
  3.body: The actual text content of the Reddit comment.
  4.downs: Number of downvotes the comment received.
  5.created_utc: The UTC timestamp when the comment was created.
  6.score: Net popularity score of the comment (upvotes − downvotes).
  7.author: Username of the person who wrote the comment.
  8.distinguished: Indicates if the comment was made by a moderator or admin.
  9.id: Unique identifier for each comment.
- **Target Variable:** Score. This score represents the net popularity, calculated as the number of upvotes minus the number of downvotes a post or comment has received

Architecture Diagram



Methodology

1. **Importing Libraries:**
   Imported essential Python libraries such as pandas, numpy, and machine learning modules from sklearn including linear regression, random forest, gradient boosting, preprocessing, and metrics.

2. **Data Loading:**
   Loaded a structured dataset containing features like ups, downs, body_length, post_hour, and a target variable representing post popularity.

3. **Data Cleaning:**
   Handled missing values by filling them with mean values of respective columns.

4. **Feature Engineering:**

   - Created interaction and transformed features (e.g., ups_downs_product, ups_squared, body_hour_interaction).

   - Applied logarithmic transformations (ups_log, downs_log).

   - Generated ratio features like ups_to_downs to capture engagement patterns.

5. **Feature Scaling:**
   Normalized features and target variable using scaling techniques to improve model performance.

6. **Model Training:**
   Trained multiple regression models including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor.

7. **Model Evaluation:**
   Evaluated models using **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)** on the validation set.

8. **Feature Importance Analysis:**
   Identified and analyzed which engineered features contributed most to predicting post popularity.

Results & Evaluation

Key Results

1. **Model Performance**
   Three regression models were trained to predict Reddit post popularity:
   - **Linear Regression**
   - **Random Forest Regressor**
   - **Gradient Boosting Regressor**

   After training and evaluation, the models were compared using **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)**.
   - **Linear Regression** performed moderately, capturing basic linear relationships but missing nonlinear interactions.
   - **Random Forest** and **Gradient Boosting** performed better, capturing nonlinear effects and interactions among features.
   - Random Forest and Gradient Boosting showed lower MSE and MAE, indicating better predictive performance.

2. **Feature Importance Analysis**: Analysis of feature contributions revealed the most influential factors for post popularity:
3. **Interpretation:**
   - Longer posts tend to get higher engagement.
   - Raw upvotes and downvotes strongly correlate with overall popularity.
   - The time of posting also affects engagement, although to a lesser extent.

Evaluation Metrics Used

1. **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual popularity scores. Lower values indicate better model performance.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2$$

2. **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual popularity scores. Lower values indicate better predictive accuracy.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \quad \mid y_i - \hat{y}_i \mid$$

Conclusion

In this project, we successfully developed regression models to predict the popularity of Reddit posts using features such as upvotes, downvotes, post length, and posting time. By applying feature engineering and scaling, we enhanced the models ability to capture interactions and non-linear relationships within the data. Among the models tested, ensemble methods like Random Forest and Gradient Boosting outperformed Linear Regression, achieving lower mean squared error and mean absolute error, indicating more accurate predictions. Feature importance analysis revealed that post length, upvotes, downvotes, and posting hour are the most influential factors affecting popularity, aligning with intuitive expectations about user engagement. Overall, this project demonstrates that machine learning can effectively model social media popularity using metadata features, and it highlights the value of preprocessing, feature engineering, and model selection in building accurate and interpretable predictive models. Integrating text-based features in future work could further improve performance and provide deeper insights into what drives post engagement.